Imperial College London

Department of Life Science

Time Lag Is a Worthwhile Parameter in Mechanistic Modelling of Bacterial Growth

Author:
Rui Zhang

Supervisor:
Samraat Pawar

Date: 1/21/2021
Word Count: 2448

Abstract

There are a lot of different models fitting population growth which can be applied to the analysis of ecosystem dynamics and emergent functional characteristics. This project focuses on cubic polynomial model, Logistic model and Gompertz model, trying to find out which model fits real microbes population growth data the best. After using R to fit models and calculating Akaike information criterion, Schwarz criterion and the residual sum of squares, Gompertz model shows the best fitting performance among these 3 models. The parameter time lag is worthy to being added into microbial growth model considering both fit goodness and model complexity.

1 Introduction

Population growth equation is necessary in dynamic models to describe the development of the ecosystem and interactions between the population and environment (Gamito 1998). Carrying capacity, population growth rate and metabolic rate of population can all be affected by environment and then results in changes of population dynamics and ecosystem dynamics (Bernhardt et al. 2018). Totally comprehension of population growth model is necessary to estimate extended domains, such as marine fisheries productivity (Sarker & Wiltshire 2017), disease transmission (Aldila & Seno 2019), even the direction of evolution (Hendry & Green 2017).

There are also many different population growth models, involving phenomenological and mechanistic models. Cubic polynomial model is the simplest phenomenological mathematical linear model, with 4 parameters, to describe population growth. For mechanistic models, a classical, somewhat mechanistic model is the logistic equation with 3 parameters (Peleg & Shetty 1997). However this model failed to match lag phase of population growth. To capture the lag phase, more complicated bacterial growth models have been designed, one of which is the modified Gompertz model with 4 parameters (Zwietering et al. 1990), which is the most frequently used model

in the research to fit population growth. When comparing these models, a 4 parameters totally phenomenological model, a 3 parameter mechanistic model and a 4 parameters mechanistic model, Zwietering used t test and F test to compare models' statistical sufficient to describe population growth and their ease of use (Zwietering et al. 1990).

Now there are some more scientific model selection criterion to answer the question whether it is worthwhile to add a new parameter to improve the precise of model fitting. The residual sum of squares, aka R^2, is a naive method to measure fit. But only considering fit is not enough. The Akaike information criterion (AIC) (Burnham 2002) calculates the Kullback-Leibler information lost during approximation full reality with the fitted model, which contains a bias correction factor with respect to model complexity. Schwarz criterion, also named Bayesian information criterion (BIC), is another model selection criterion considering both fit and complexity. However, it includes a penalty term dependent on dataset size and tends to choose simpler models (Johnson & Omland 2004). Generally, AIC is prefered by researchers because of based on Kullback-Leibler information.

Usually, people favor simpler models due to Ockham's razor (aka parsimony principle) if models describe the same pattern adequately (Peleg & Corradini 2011). To find out whether time lag is a worthwhile parameter to be introduced in mechanistic model fitting and whether a more complicated form of equation works better than totally phenomenological linear model, I fit cubic polynomial, Logistic and Gompertz models to a large amount of microbes population growth experiments data with R. Bacteria cultivation is easy and standardized, and it is convenient and fast to get population growth data. Furthermore, bacteria appears to have a large biodiversity and large habitat diversity which plays an important role in human health, food safty, fermentation engineering, environment and so on. R is a platform that is mature and convenient to fit ecological model and analyze data, based on numerous support packages, help documents, simple installation

and convenient high-level syntax (Bolker et al. 2013). By means of comparing statistical measure of model fitting, we can discover which model is the most suitable to describe microbes population growth, especially bacterial growth.

2 Materials & Methods

2.1 Materials

The dataset contains 4388 records, collected from 10 lab experiments data across the world, containing the change in biomass or population of microbes over time. The species in these microbes data include bacteria and mobile phytoplankton, which play an important role in food safty and environment. Time range of the dataset is also quite large, varying from 1962 to 2018, which could guarantee the universality of model comparison conclusions.

2.2 Methods

2.2.1 Computing tools

Data processing, model fitting and results visualization are performed with R due to its excellent graphing capability and numerous packages for statistical analysis and data handling, including reshape2, minpack.lm and ggplot2 packages used in this program. Python is an ideal tool to build an automated workflow to analyze data with the subprocess module, especially useful when the program involves many different languages. Therefore, I use shell commands in Python script to run R scripts and compile LATEX to create the report, which guarantees this program is fully reproducible.

2.2.2 Data management

In this program, two main variables of interest are time and population. Therefore, I create a new ID with all of the other variables, with which I divide the total dataset into 285 subdatasets. Then I delete negative time

and population biomass, during which I find an unreasonable dataset. This subdataset contains 29 records but 19 biomass values are negative, so I delete this dataset totally. Finally there are 284 datasets, containing 4284 records to fit cubic polynomial model, Logistic model and Gompertz model.

2.2.3 Model fitting

There I choose cubic polynomial model, Logistic model and Gompertz model to fit all of 284 datasets. Because of subsequent AIC and BIC comparison, the formation of biomass or population should be the same. To guarantee all of these equation have the same response variable, I transform cubic polynomial model and Logistic model into log scale.

The equation of cubic polynomial model is:

$$log(N_t) = log(a + bt + ct^2 + dt^3) \tag{1}$$

The equation of Logistic model is:

$$log(N_t) = log(\frac{N_0 K e^{rt}}{K + N_0(e^{rt} - 1)}) \tag{2}$$

The equation of Gompertz model is:

$$log(N_t) = N_0 + (N_{max} - N_0)e^{-e^{\frac{r_{max}e}{(N_{max}-N_0)log(10)}^{t_{lag}^{-t}}+1}} \tag{3}$$

a, b, c, d are all parameters without biological meaning. N_t is biomass or population value at time t, N_0 is biomass or population value at initial condition, N_{max} and K are carrying capacity, r and r_{max} are maximum growth rate and t_{lag} is time lag which is the duration of delay before the population grow exponentially.

Before model fitting, I transform biomass and population data into log scale and then fit experiments data to models. For the simple linear phenomenological model, cubic polynomial model, I use R to fit linear model directly. For Logistic and Gompertz nonlinear models, I need to calculate and choose

start value of parameters. N_0 start value is the minimum of biomass or population. N_{max} and K are same, whose start value is the maximum of biomass or population. Start value of r is calculated by linear model, choosing the middle 70 percents of biomass or population value range, drawing a straight line and setting the value of slope as start value. t_{lag} starting value is set as the time of the largest second order derivative of population growth point. Furthermore, I sample start values 400 times to find out the best start value combination.

After model fitting, I calculate AIC, BIC and R^2 of these model fittings. I also calculate AICc, which is the second order derivative of AIC and correct small sample size. When size/40 < number of parameters and size > 5, I calculate AICc as a supplementary criteria of model selection. Here is the equation of AICc (k is the number of free parameters and n is the sample size):

$$AIC_c = AIC + \frac{2k^2 + 2k}{n - k - 1} \tag{4}$$

2.2.4 Plotting and analysis

I draw line charts of AIC, BIC and R^2 which give a direct discription on 3 models' fitting performance in bacterial growth data. Moreover, the smaller AIC and BIC are and the bigger R^2 is, the better model performs. Therefore I calculate the best performing model propotion in AIC, BIC and R^2 criteria and visualize them as pie charts. Although only part of datasets meet the condition of using AICc, I analyze it and summarize it with the other 3 model comparison criteria in a table.

3 Results

Finally, 282 of 284 datasets are fitted successfully in all of cubic polynomial model, Logistic model and Gompertz model. Also, all of the model fitting lines are drawn in the point graph of every dataset. Here is an example

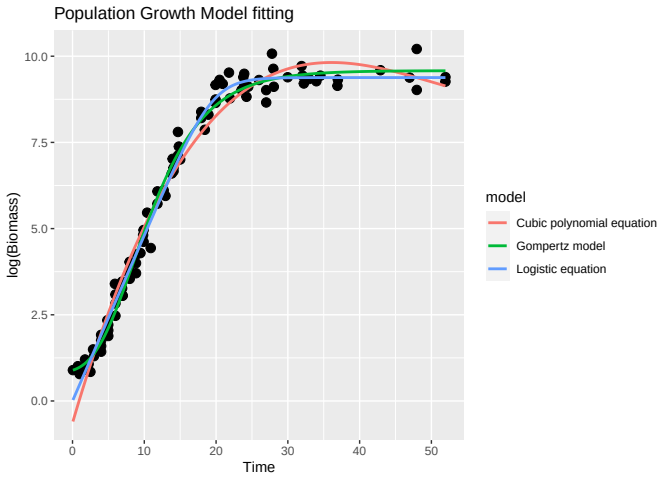

Figure 1: An example of 3 models fitting to a dataset

of the model fitting situation to the last dataset (Figure 1). This dataset contains 89 records, which is a relative big sample size among 284 datasets, and the population growth curve is also reasonable. It is obvious that every model shows a good fit to the growth of microbes population, but only Gompertz model catches the lag phase of microbe growth. Therefore, it seems like Gompertz model is the best model to fit this dataset. However, there are some more scientific and reasonable model selection criterias considering fit and complexity and giving a specific number to show model fitting performance, such as AIC, AICc, BIC and R^2. Table 1 demonstrates exact number of datasets that every model shows the best performance in AIC, AICc, BIC and R^2.

Table 1: The best fitting model under 4 criteria

Model	AIC	AICc	BIC	R^2
Gompertz	172	127	169	192
Logistic	54	112	59	29
Cubic	56	25	54	61

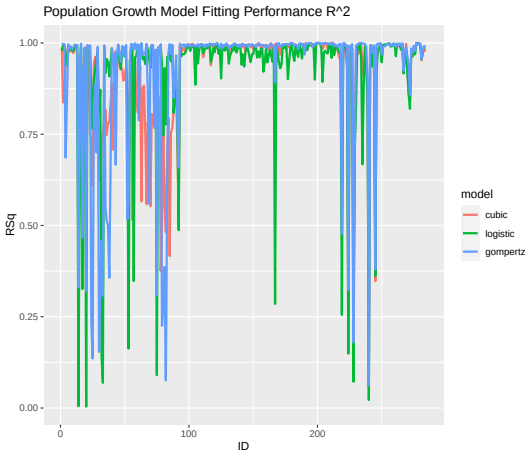

Figure 2: Model fitting performance R^2

3.1 R^2

R^2 is the residual sum of squares for a model, which is the simplest criterion to show the quality of fit. The more R^2 close to 1, the better model fits experiments data. As shown in Figure 2, cubic model shows general large R^2, while both of Logistic model and Gompertz model appears to have some quite low R^2. Calculated from Table 1, there are 68.1% datasets fitted best by Gompertz model, 21.6% datasets fitted best by cubic polynomial model and 10.3% datasets fitted best by Logistic model. However, maximizing R^2 doesn't consider model complexity, neglecting the parsimony principle.

Figure 3: Model fitting performance AIC

3.2 AIC & AICc

The Akaike information criterion (AIC) is a model selection criterion that considers both fit and complexity among multiple models. When the sample size is small, AICc should be used to correct the bias. Gompertz model's AIC shows an overall relative low value compared with the other 2 models given Figure 3 line chart. In further best fitting model analysis, Gompertz model, cubic polynomial model and Logistic model fitting real microbes population growth data best accounts for 61%, 19.9% and 19.1% of 284 datasets respectively. However, when calculating AICc, there are 18 datasets' size less than or equal to 5, which can not be calculated AICc value. In 264 datasets, best fitting model propotion changes a lot, Gompertz model decreasing to 48.1%, cubic polynomial model decreasing to 9.5% and Logistic model increasing to 42.4%. But Gompertz model still accounts for the largest propotion of best fitting and shows general lower AICc value in Figure 4.

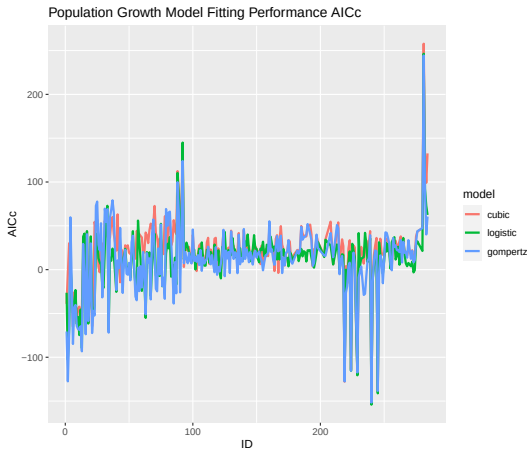

Figure 4: Model fitting performance AICc

3.3 BIC

BIC is a model selection criterion taking fit ,complexity and sample size into account. It prefers simpler models, especially sample size increasing. Three models' fitting performance in Bayesian information criterion (Figure 5) is similar with what AIC interprets, Gompertz model showing a globle lower BIC. The best performing model propotion in BIC is 59.9% for Gompertz model, 19.1% for cubic polynomial model and 20.9% for Logistic model.

4 Discussion

Population growth is a complex system affected by massive factors. What is a good method to deal with such a complicated system is setting up mathematical models. 3 criteria of a mathematical model is generality, realism and precision (LEVINS 1966). When establishing models, we should balance these 3 criteria according to our research objective. Strictly speaking, three models in this program are all empirical models instead of totally mechanistic models. The purpose of these models is to describe and quan-

Figure 5: Model fitting performance BIC

tify experimentally observed patterns, but not explain why this particular pattern emerges (Peleg & Corradini 2011). Therefore, the mathematical convenience and good fit performance matter. Here I use AIC, BIC and R^2 to show statistical measure of model fitting, and try to answer whether time lag is worthy of introduced in microbes population growth model.

In the data management, I deleted a dataset with too high propotion of negative biomass. Therefore, the other 284 datasets are fitted by cubic polynomial model, Logistic model and Gompertz model after deleted negative biomass and time. Ultimately, 282 datasets are successfully fitted in all of these 3 models, and the fitting lines seem rational and well-fitted to describe microbes population growth. R^2 is used to show the goodness of fit, while AIC, AICc and BIC are used as model selection criteria, considering both fit and model complexity. Cubic polynomial model seems to fit microbes experiments data well with general high R^2 value while Gompertz model shows the highest R^2 value in 192 datasets. For model selection, Gompertz model shows the lowest AIC and BIC value in majority of 282

datasets, which means although BIC favor simpler model and have correct sample size bias, Gompertz model still performs better than the other 2 models. When considering AICc, it is the same case. In conclusion, Gompertz model is the best of these 3 models to describe microbes population growth.

The success of Gompertz model demonstrates at least 2 enlightenments. Firstly, mechanistic model is better than totally phenomenological linear model when the number of parameters are the same. Time lag is a worthwhile parameter to be added into microbes population growth data which can improve model fitting performance significantly. The biological reason of this phenomenon may be the transition behavior of microbes at the end of lag phase of population growth process (Verhulst et al. 2011). Although the value of parameter time lag might not be equal to the biological meaning of lag phase duration.

Besides, there are many other population growth model not involved in this program. Such as Baranyi model (BARANYI et al. 1993), three-phase model (Buchanan et al. 1997), totally and truly mechanistic model on cells level (Peleg & Corradini 2011) and so on. Baranyi model is also a widely used model to describe population growth. However, instead of adding time lag parameter, it added the initial physiological state of cells h_0 compared with Gompertz model. Furthermore, this model used an equation $h_0 = t_{lag} r_{max}$, appearing in the solution of rate equation, to calculate time lag. The ralative high correlation between r_{max} and h_0 results in difficulty of estimating the parameters value of Baranyi model (Grijspeerdt & Vanrolleghem 1999). But a more district experiments design can guarantee the precise of parameter estimate. Also the benefits of introducing h_0 in model should be calculated in the future.

Whitting specified models into 3 levels (Whiting 1995). All models mentioned above are all level one model that only describe changes of microbial

numbers versus time. Level 2 models demonstrates effect of environment on parameters in level 1. What level 3 models do is combining level 1 and level 2 models and calculating microbial behavior under changed environment condition. We should choose suitable models according to our objective. Such as cubic polynomial model can catch death phase, that may be the reason why in some datasets cubic model shows the best fit. If we need to predict different temperature population growth (level 3 models), Gompertz model has some shortcomes and Baranyi model is the better choice of level 1 model (Peleg & Corradini 2011, Silva et al. 2018). When using models, we ought to keep in mind that an appropriate model to the question is more important than model fitting performance. The validation of a model is that it generates good testable hypotheses relevent to crucial problems.

In summary, Gompertz model fits microbes population growth expirical datasets best because of catching lag phase of microbes growth which is consistent with microbes transition behavior observed in experiments. In addition, time lag is worthy to be added into fitting model after balancing fit goodness and model complexity.

References

Aldila, D. & Seno, H. (2019), 'A population dynamics model of mosquito-borne disease transmission, focusing on mosquitoes' biased distribution and mosquito repellent use', *Bulletin of mathematical biology* **81**(12), 4977–5008.

BARANYI, J., ROBERTS, T. A. & MCCLURE, P. (1993), 'A non-autonomous differential equation to model bacterial growth', *Food microbiology* **10**(1), 43–59.

Bernhardt, J. R., Sunday, J. M. & O'Connor, M. I. (2018), 'Metabolic theory and the temperature-size rule explain the temperature dependence of population carrying capacity', *The American naturalist* **192**(6), 687–697.

Bolker, B. M., Gardner, B., Maunder, M., Berg, C. W., Brooks, M., Comita, L., Crone, E., Cubaynes, S., Davies, T., de Valpine, P., Ford, J., Gimenez, O., Kéry, M., Kim, E. J., Lennert-Cody, C., Magnusson, A., Martell, S., Nash, J., Nielsen, A., Regetz, J., Skaug, H. & Zipkin, E. (2013), 'Strategies for fitting nonlinear ecological models in r, ad model builder, and bugs', *Methods in Ecology and Evolution* **4**(6), 501–512. **URL:** *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12044*

Buchanan, R., Whiting, R. & Damert, W. (1997), 'When is simple good enough: a comparison of the gompertz, baranyi, and three-phase linear models for fitting bacterial growth curves', *Food microbiology* **14**(4), 313–326.

Burnham, K. P. (2002), *Model selection and multi-model inference : a practical information-theoretic approach*, 2nd ed. edn, Springer, New York ; London.

Gamito, S. (1998), 'Growth models and their use in ecological modelling: an application to a fish population', *Ecological Modelling* **113**(1), 83 – 94. **URL:** *http://www.sciencedirect.com/science/article/pii/S0304380098001367*

Grijspeerdt, K. & Vanrolleghem, P. (1999), 'Estimating the parameters of the baranyi model for bacterial growth', *Food microbiology* **16**(6), 593–605.

Hendry, A. P. & Green, D. M. (2017), 'Eco-evolutionary dynamics in cold blood', *Copeia* **105**(3), 441–450.

Johnson, J. B. & Omland, K. S. (2004), 'Model selection in ecology and evolution', *Trends in Ecology and Evolution* **19**(2), 101 – 108. **URL:** *http://www.sciencedirect.com/science/article/pii/S0169534703003458*

LEVINS, R. (1966), 'The strategy of model building in population biology', *American scientist* **54**(4), 421–431.

Peleg, M. & Corradini, M. G. (2011), 'Microbial growth curves: What the models tell us and what they cannot', *Critical Reviews in Food Science and Nutrition* **51**(10), 917–945. PMID: 21955092.
URL: *https://doi.org/10.1080/10408398.2011.570463*

Peleg, M. & Shetty, D. K. (1997), 'Modeling microbial populations with the original and modified versions of the continuous and discrete logistic equations', *Critical Reviews in Food Science and Nutrition* **37**(5), 471–490. PMID: 9315435.
URL: *https://doi.org/10.1080/10408399709527785*

Sarker, S. & Wiltshire, K. H. (2017), 'Phytoplankton carrying capacity: Is this a viable concept for coastal seas?', *Ocean and Coastal Management* **148**, 1 – 8.
URL: *http://www.sciencedirect.com/science/article/pii/S0964569117301977*

Silva, A. P. R. d., Longhi, D. A. & Dalcanton, F. (2018), 'Modelling the growth of lactic acid bacteria at different temperatures', *Brazilian Archives of Biology and Technology* **61**.
URL: *https://doi.org/10.1590/1678-4324-2018160159*

Verhulst, A., Cappuyns, A., Van Derlinden, E., Bernaerts, K. & Van Impe, J. (2011), 'Analysis of the lag phase to exponential growth transition by incorporating inoculum characteristics', *Food microbiology* **28**(4), 656–666.

Whiting, R. C. (1995), 'Microbial modeling in foods', *Critical Reviews in Food Science and Nutrition* **35**(6), 467–494.
URL: *https://doi.org/10.1080/10408399509527711*

Zwietering, M. H., Jongenburger, I., Rombouts, F. M. & van 't Riet, K. (1990), 'Modeling of the bacterial growth curve', *Applied and Environmental Microbiology* **56**(6), 1875–1881.